DOCUMENT RESUME

ED 116 147 CS 002 311

AUTHOR Satz, Paul; Friel, Janette
TITLE The Predictive Validity of an Abbreviated Screening Battery: A Preliminary Cross Validation Study.
PUB DATE 75
NOTE 20p.; Paper presented at the Annual Meeting of the American Psychological Assn. (83rd, Chicago, Illinois, August 30-September 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 Plus Postage
DESCRIPTORS Educational Research; Evaluation Methods; Kindergarten Children; Longitudinal Studies; *Measurement Instruments; Predictive Measurement; *Predictive Validity; Primary Education; *Reading Difficulty; *Reading Skills; *Reading Tests

ABSTRACT

This study determines whether an abbreviated test battery, administered in September, could predict achievement ratings at the end of kindergarten in June of a group of kindergarten children in an elementary school. An additional purpose was to institute a prevention program on a random sample of predicted high-risk children in this group and to evaluate the test outcomes despite the possible ameliorative effects of treatment. The sample consisted of 28 black students and 104 white students who entered kindergarten in September 1974. The results provide additional support for the predictive validity of this abbreviated screening battery. The major reservation concerning the results of this study is the tentative, if not premature, state of the achievement criterion. (RB)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED116147

The Predictive Validity of an Abbreviated Screening Battery: A Preliminary Cross Validation Study[1]

Paul Satz and Janette Friel
University of Florida

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
Paul Satz
Janette Friel
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

CS 002 311

An abbreviated behavioral screening battery has been developed which is designed to detect those children (white males) during kindergarten who in four years will become severely disabled or superior readers (Satz & Friel, 1973; Satz & Friel, 1974; Satz, Friel & Rudegeair, 1975). Standardization of the test battery was based on the total population of white boys (N=497) who started kindergarten in the fall of 1970 in Alachua County, Florida. Twenty elementary schools (14 urban, 6 rural) comprised this population. The predictive tests (N=8), given during early kindergarten, were later validated against independent reading criteria at the end of Grades 1 (1972), 2 (1973) and 3 (1974).[2] Because of careful tracking procedures, the follow-up validation studies were based on approximately 95 percent of the original population. A separate cross validation study was recently completed based on a sample of white boys (N=181) from five of the urban schools (Alachua County) who began kindergarten in 1971 and whose reading scores were assessed three years later at the end of Grade 2 (1974). The children in this sample were predicted on the basis of their abbreviated test battery scores (kindergarten) and the discriminant function weights derived from the original standardization population at the end of Grade 2.

The results of the preceding validation studies, based on discriminant function analyses, revealed that the tests given during kindergarten consistently identified over 90 percent of the children destined to become severely disabled or superior readers in later years. Stepwise regression analyses also revealed that a small number of tests consistently ranked highest in predicting reading outcomes in later grades (1-3). These tests (Finger Localization, Alphabet Recitation, Recognition-Discrimination) loaded on a primary factor which was labeled a measure of sensori-perceptual-motor-

mnemonic ability. This factor is felt to tap those skills which are in primary ascendancy during preschool years (kindergarten) and which are postulated to be crucial to the early phases of reading (Satz & Van Nostrand, 1972).[3] A major review of the theory guiding this research and the results are reported in a recent chapter by Satz, Friel & Rudegeair (1975). The results, in summary, lend support to the validity and utility of an early detection or 'warning system' that could be administered economically before the child begins formal reading--at a time when his central nervous system may be more plastic and responsive to change--and at a time when he is less subject to the shattering effects of repeated academic failure.

Despite the promising nature of this early detection research, more cross validation work is needed to ensure that truly high risk children are identified for placement into early prevention programs. For example, if the tests have a high number of false positives, then serious risks of mislabeling could occur which would incorrectly place potentially normal readers into early treatment programs. Similarly, if the false negative rates are high, then unnecessary costs may result by instituting treatment programs which exclude the majority of high risk children. An additional need for further cross validation study is to determine whether the battery is applicable to girls (white & black) and to black boys.

The purpose of the present study was to determine whether the abbreviated test battery, given in September, could predict achievement ratings at the end of kindergarten (June) on a new group of kindergarten children (boys, girls, black, white) in an entire elementary school. An additional purpose was to institute a prevention program on a random sample of <u>predicted</u> high risk children in this group and to evaluate the test outcomes de-

spite the possible ameliorative effects of treatment. This design was felt to provide a more powerful test of the predictive efficiency of the abbreviated battery.

Method

Subjects

The sample consisted of 28 black Ss (13 boys, 15 girls) and 104 white Ss (54 boys, 50 girls) for a total of 132 Ss. This group represented all those Ss who entered kindergarten in September, 1974 at Stephen Foster School (mean age=54.1 months). Additional age and SES information is presented in Table 1 for outcome achievement groups at the end of kindergarten.

Predictor Tests

The abbreviated test battery consisted of seven tests and one non-test variable (SES). The seven tests were: Finger localization, Alphabet Recitation, Recognition-Discrimination, Peabody IQ, Beery (VMI), Wepman Auditory Discrimination and Dichotic-Listening (Total R+L Score).[4] A recent factor analysis of this battery, based on the original standardization and cross validation groups (white male, N=678), revealed two main factors using an oblique solution.[5] Factor I, which comprised most of the verbal tests, included Peabody IQ, Alphabet Recitation, Auditory Discrimination and Dichotic-Listening. Factor II, which comprised most of the nonverbal perceptual tests, included Recognition-Discrimination, Beery (VMI) and Finger Localization. These factor loadings are different than those reported for the standard 14 test battery (Satz & Friel, 1973).

Test Procedure

The tests were given during two weeks in September in office space provided by the elementary school. Intervention procedures were carried

out in a mobile laboratory parked on school grounds. All testing was administered individually by the research staff and took approximately 50 minutes per child. At the conclusion of the test administration, all results were analysed via two computer programs (DSCRIM[6] and CLASIFY) which utilized the lambda weights derived from the pooled standardization and cross validation population of white boys (N=633). Based on this analysis, 44 children were identified as severe high risk (‡) and 28 of them were randomly assigned to two treatment groups for the duration of the school year. The remaining Ss were placed into a non-treatment group (N=16).[7] This random assignment yielded approximately equal numbers of children (by race and sex) in each of the three groups. To prevent individual labeling of children, all test information was withheld from the teachers during the school year. As a further control, selected children from the other predicted groups (low risk) were periodically brought to the trailer during the year for additional research study.

Achievement Criterion Ratings

The achievement ratings were obtained at the end of the school year (May-June) for each child. The ratings were made by the individual classroom teachers who had taught the children during the school year. An overall achievement rather than reading criterion was used because of difficulty in assessing reading competency at this age. A ten-point interval scale was used to assign the children into one of four different achievement groups: Severe (0-4), Mild (5-6), Average (7-9) and Superior (10). Criterion information was available for 128 of the original children, or 97% of the sample: Severe (N=12), Mild (N=33), Average (N=63) and Superior (N=20). This criterion evaluation, while admittedly tentative, if not premature, has nevertheless been shown to hold up in later years when more objective

reading measures are available (Satz, Friel & Rudegeair, 1975).

Results

Cross Validation

Means and Standard Deviations. The means and standard deviations of the tests given in September are presented in Table 1 for each of the four outcome achievement groups (June). Inspection of this table reveals an increasing level of performance across tests as one proceeds from the Severe to Superior groups. For example, mean Peabody IQ ranged from 77.7 in the Severe group to 113.6 in the Superior group. Similarly, performance on the Beery (VMI) ranged from a mean score of 49.6 months in the Severe group to 67.4 months in the Superior group. This means that the Severe group was lagging almost 13 months behind their chronological age (62.5) whereas the Superior group was advanced almost three and a half months beyond their chronological age (64.1).

Insert Table 1 about here

Classification. Subjects were then classified into the predicted achievement groups based on the tests given in September plus the weights derived from the original standardization and cross validation groups. The results, for the final outcome achievement groups (June), are presented in Table 2 for a 2x4 matrix. That is, the composite test predictions (September) are reduced to high risk ($\overset{+}{+}$ & +) and low risk ($\overset{+}{-}$ & -) signs and are represented by rows whereas the achievement outcomes (June) are represented by columns. Inspection of this table reveals that the tests correctly predicted 100% of the Severe and Superior groups, while misclassifying 21% of the Mild group (N= 7) and 41% of the Average group (n=26). The overall hit-

rate was 95/128=74% for this cross validation sample.

Insert Table 2 about here

The classification outcomes for a 4x4 matrix are presented in Table 3. This table represents a more meaningful presentation of results because it reveals the outcomes for the severe high risk predictions (‡) which formed the decisional basis for the treatment programs (experimental and control) in September (Grade K). The only difference in this table is that the composite predictions are reduced to four levels (rows) which does not alter the overall hit-rate. Inspection of this table reveals that these severe high risk indicators (‡) detected 100% of the children who at the end of kindergarten fell in the Severe group and 58% of those who fell in the Mild group. Although it misclassified 13 children (20%) who fell in the Average group at the end of kindergarten, it did not misclassify any children who later fell in the Superior group.

Insert Table 3 about here

In other words, there were only 13 misclassification errors using the severe high risk signs, all of which were confined to the Average group. No children with these risk signs ended up in the Superior group. Moreover, when these children with severe risk signs were examined for treatment group assignment (experimental vs. control) it was found that eight of the 13 misclassified children were in treatment groups, which reduces the predictive error to only five children--again, none of whom ended up in the Superior group![8]

rate to 84.4%.

Discussion

The preceding results provide additional cross validational support for the predictive efficiency of this abbreviated screening battery. The tests given at the beginning of Grade K, based on the weights derived from the standardization and cross validation groups (white males), correctly predicted the achievement outcomes of a new sample of children (boys, girls, black, white) at the end of Grade K--particularly those destined to extremes in the achievement distribution. Furthermore, these predictions held up despite the fact that the sample varied in terms of race and sex and the fact that a treatment program was instituted for a majority of the severe high risk children. In fact, it was shown that the validity of the severe high risk test sign (‡) was extremely high; it detected 100% of the children who fell in the Severe group at the end of kindergarten while misclassifying only 20% of the children (N=13) who fell in the Average group (Table 3). However, this test sign (‡) misclassified no children who, at the end of kindergarten, fell in the Superior group. Moreover, of the 13 high risk children (20%) who fell in the Average group at the end of kindergarten, eight of them were involved in individual treatment programs throughout the entire year which suggests that treatment per se may have altered the high risk signs (false positives) seen at the beginning of kindergarten. These findings, in summary, reduce the risk associated with false positive decision errors without increasing the false negative errors; i.e., 100% of the Severe cases were correctly detected.

The major reservation with the present results concerns the tentative, if not premature state of the achievement criterion. The interval of nine months between tests and criterion probes, particularly during kindergarten, provides only tentative information, at best,

on the validity of the current achievement ratings. Although predictions of learning ability based upon teacher judgments have been shown to be surprisingly accurate (Austin & Morrison, 1963; Kermoian, 1962; Feshbach, Adelman & Fuller, 1974), the fear of mislabeling may increase the incidence of false negative errors, particularly for the severely high risk child. This problem was seen recently in an unpublished study which compared teacher predictions (end of Grade K) and test predictions (beginning Grade K) to reading outcomes at the end of Grade 2. The study was based on the third year follow-up of the population of white boys who began kindergarten in Alachua County, Florida (1970) and who represent the standardization group for the present abbreviated battery (Satz, Friel & Rudegeair, 1975). The results showed that while the overall accuracy of kindergarten teacher predictions was as high as the tests (approximately 80%), the detection of the severely high risk child (Grade 2) was much lower when predicted by the teachers. The teachers identified only 19% of these children whereas the tests detected 75% of them. In other words, the overall teacher predictions were spuriously inflated by 'good outcome' forecasts when the base rates favored such outcomes (by 4:1). However, when they predicted severe outcomes (which was rare), their accuracy was extremely high (approximately 90%).

This unpublished finding is relevant to the current kindergarten achievement criterion. It suggests that those children designated Severe may in fact turn out to be so, but not those classified as Average, many of whom in later years may fall in the Mild to Severe achievement groups. If so, it would further lower the false positive rate of the abbreviated test battery, especially when predictions are followed-up in later years. This is indeed what happened when the test predictions (Grade K) were evaluated

against achievement criteria at the end of Grade K (Satz & Friel, 1973), Grade 1 (Satz, Friel, 1974), Grade 2 (Satz, Friel & Goebel, 1975) and Grade 3 (Satz, Friel & Rudegeair, 1975). In later years, the predictive accuracy of the tests (Grade K) increased with incremental reductions in the false positive rate--again, due presumably to increased validity in the criterion achievement measures.

This problem regarding preliminary criterion specification (at Grade K) may also explain, in part, the change in the discriminative ranking of the tests. It was shown that the four tests which ranked highest in the stepwise regression analysis were Peabody IQ, Alphabet Recitation, Dichotic Listening and SES (three of which loaded on the verbal factor). This discriminative ranking, however, contrasted with the triad of Finger Localization, Alphabet Recitation, and Recognition-Discrimination which consistently ranked highest in later years (Grades 1-3) and which loaded on the sensori-perceptual-motor factor. Although the change in discriminative ranking may in part reflect changes in the current school sample (i.e., girls & blacks), the change more likely reflects the nature of the kindergarten criterion. The reason is that approximately the same discriminative ranking occurred in the standardization population at the end of Grade K (Satz & Friel, 1973) [ref. footnote 3].

The preceding explanations, of course, must ultimately rest on the follow-up evaluations at the end of Grades 1 and 2 with this cross validation sample.[9] If confirmed, they will provide additional support for the utility of an early warning system that could be administered economically before the child begins formal reading--at a time when the central nervous system may be more plastic and responsive to change--and at a time when the child is less subject to the shattering effects of repeated academic failure.

The ultimate task for education is to prevent the needless suffering that results when a system fails to develop a valid early screening program for its high risk children.

References

Austin, M. C. & Morrison, C. The first R: The Harvard Report on reading in elementary schools. New York: Macmillan, 1963.

Feshbach, S., Adelman, H. & Fuller, W. W. Early identification of children with high risk of reading failure. Journal of Learning Disabilities, 1974, 7, 49-54.

Kermoian, S. B. Teacher appraisal of first grade readiness. Elementary English, 1962, 39, 196-201.

Satz, P., & Friel, J. Some predictive antecedents of specific learning disability: A preliminary one-year follow-up. In P. Satz and J. Ross (Eds.), The disabled learner: Early detection and intervention. Rotterdam,The Netherlands: Rotterdam University Press, 1973. Pp.79-98.

Satz, P. & Friel, J. Some predictive antecedents of specific reading disability: A preliminary two-year follow-up. Journal of Learning Disabilities, 1974, 7, 437-444.

Satz, P., Friel J. & Goebel, R. Some Predictive antecedents of specific reading disability: A three-year follow-up. Bulletin of the Orton Society, In press, 1975.

Satz, P., Friel, J. & Rudegeair, F. Some predictive antecedents of specfic reading disability: A two- three- and four-year follow-up. In J. T. Guthrie (Eds.), Aspects of Reading Acquisition, Baltimore: Johns Hopkins Press, In press, 1975.

Satz, P. & Van Nostrand, E. K. Developmental dyslexia: An evaluation of a theory. In P. Satz and J. Ross (Eds.), The disabled learner: Early detection and intervention. Rotterdam, The Netherlands: Rotterdam University Press, 1972. Pp. 121-148.

Footnotes

[1]This research was supported in part by funds from the National Institutes of Health (NS08208) and the National Institutes of Mental Health (MH19415).

[2]The original screening battery consisted of 14 tests (Satz, Friel & Rudegeair, 1975).

[3]This discriminative ranking however, was not observed during the first follow-up evaluation of achievement outcomes at the end of kindergarten (Satz & Friel, 1973). The ranking in this preliminary study was as follows: Finger Localization, SES, Dichotic Listening (Total) and Peabody IQ.

[4]Consult Satz, Friel & Rudegeair (1975) and Satz & Friel (1973) for additional information concerning description and selection of the test battery.

[5]The factor analysis was computed by Jack Fletcher.

[6]Written by D. J. Veldman, University of Texas, 1967; modified by R. A. Goebel. Statistical analyses performed on IBM System/370-165, Northeast Regional Data Center, Gainesville, Florida.

[7]The treatment program and rationale is not relevant to the present study but will be discussed in a later study after follow-up evaluations are made in Grade 1.

[8]It was also found that six of the 12 children in the Severe group and 11 of the 19 children in the Mild group were in treatment groups throughout the kindergarten year.

Footnotes

[9]Additional cross-cultural information on the predictive validity of this abbreviated battery will be available in the fall of 1975. The battery was administered to a representative sample of white boys (N=450) who began elementary school in the Catholic System in Perth, Australia in February, 1975. This study will also investigate whether a further abbreviation of the test battery (N=6) will predict as well as the standard abbreviation (N=8). Unpublished results have just shown that the modified abbreviation (excluding Dichotic Listening and Auditory Discrimination) predicts as well as the standard abbreviation. In fact, the modified abbreviation yielded a much lower false positive rate when applied to the current cross validation sample. If these results can be replicated on the Australian sample, they will substantially reduce the cost of administration for mass screening.

Table 1

Means and Standard Deviations of Abbreviated Test Battery (Sept., Kindergarten) for Achievement Groups (June, Kindergarten) Cross Validation Sample II[a]

Tests	Criterion Achievement Groups			
	Severe (N=12) Age=62.5[b]	Mild (N=33) Age=60.8	Average (N=63) Age=63.0	Superior (N=20) Age=64.1
1. Finger Localization	28.6 (5.2)	34.7 (7.1)	37.8 (5.8)	40.5 (3.9)
2. Alphabet Recitation	11.2 (8.7)	18.5 (6.8)	22.1 (6.7)	25.3 (2.3)
3. Recognition-Discrimination	6.5 (2.2)	7.5 (2.3)	8.7 (2.8)	10.2 (2.8)
4. Peabody IQ	77.7 (16.2)	89.2 (19.0)	103.4 (15.3)	113.6 (7.5)
5. Beery (VMI)[c]	49.6 (9.1)	53.2 (9.6)	61.6 (9.7)	67.4 (10.5)
6. Auditory-Discrimination	1.0 (0.6)	1.3 (0.4)	1.5 (0.3)	1.6 (0.3)
7. Dichotic Listening (Total)	58.7 (15.9)	70.4 (12.2)	73.4 (14.0)	80.9 (6.3)
8. Socio-economic Status	1.4 (0.5)	1.7 (0.5)	1.9 (0.3)	2.0 (0.0)

[a] Sample = all Kindergarten classes (Stephen Foster School) including girls and boys (white and black).

[b] Age in months

[c] Score in months

Table 2

Predictive Classification of Cross Validation Sample II (Sept., Grade K) into Achievement Groups (June, Grade K) based on Discriminant Function Weights (Abbreviated Battery) of Standardization Population (N=639)[a,b]

(2 x 4 Matrix)

Composite Discriminant Scores		Criterion Achievement Groups			
		Severe	Mild	Average	Superior
+	N	12	26	26	0
	%	(100)	(80)	(41)	(0)
-	N	0	7	37	20
	%	(0)	(20)	(59)	(100)
T		12	33	63	20

[a] Population = Standardization Group (Grades K-2) and Cross Validation Group I (Grades K-2).

[b] Total Hits = 95/128 = 74%.

Table 3

Predictive Classification of Cross Validation Sample II (Sept., Grade K) into Achievement Groups (June, Grade K) based on Discriminant Function Weights (Abbreviated Battery) of Standardization Population (N=639)[a]

(4 x 4 Matrix)

Composite Discriminant Scores	Criterion Achievement Groups				
	Severe	Mild	Average	Superior	T
+ +	12	19	13	0	44
+	0	7	13	0	20
-	0	5	27	10	42
- -	0	2	10	10	22
T	12	33	63	20	128

[a] Population = Standardization Group (Grades K-2) and Cross Validation Group I (Grades K-2).

Table 4

Predictive Classification of Cross Validation Sample II (Sept., Grade K) into Achievement Groups (June, Grade K) based on Discriminant Function Composite Scores (Sample II)[a]

(2 x 4 Matrix)

Composite Test Scores		Criterion Achievement Groups			
		Severe	Mild	Average	Superior
+	N	12	25	12	0
	%	(100)	(76)	(19)	(0)
-	N	0	8	51	20
	%	(0)	(24)	(81)	(100)
T		12	33	63	20

[a] Total Hits = 108/128 = 84%

Table 5

Discriminative Ranking and Cumulative Hit Classification of Abbreviated Test Battery based on Stepwise Discriminant Function Analysis of Cross Validation Sample II

Ranked Variables	Factor	Cumulative Hits (%)
1. Peabody IQ	I	75.8
2. Alphabet Recitation	I	81.3
3. Dichotic Listening (Total)	I	77.3
4. SES	-	80.5
5. Residual Tests	I-II	84.4